European Data
Grant agreement number: RI-283304

Second EUDAT Conference New Services Summary

Mark van de Sanden – SURFSARA

November 2013

TABLE OF CONTENT

LIST OF FIGURES

NEW SERVICES PARALLEL TRACK

During the 2nd EUDAT Conference held in Rome – 28-30 October 2013, one parallel track was dedicated to the New Services under discussion. Track 4 presented EUDAT work on subject areas which were earlier identified and prioritized while engaging and discussing the needs and requirements from communities and individual users. Therefore this track was named New Services and was organized in 3 separate sessions on the specific subject areas: Semantic Annotation, Dynamic Data[1] and Workflows. The reason for these topics was that we wanted to discuss the outcome of the four working groups[2] in a wider forum of possibly interested people. The first session started with an overview on the way EUDAT is collaborating with communities, users and domain experts to get an in-depth understanding on these subjects, as this is central to the way EUDAT works. This chapter provides an overview on the content presented and discussions held in the three sessions.

New Services Conclusions:

This track attracted a diverse audience from field experts to people who were interested in one of the 3 subjects being presented and discussed. The presentations gave a good overview of the results from the working group workshop held on the 25-26th September in Barcelona, community use cases to explain the specific subjects from a science domain point of view and the work conducted in EUDAT. During the different sessions there were many good discussions, in which opinions and views were shared and which will be assessed in a broader EUDAT context.

EUDAT will continue the working group discussions in the four groups to work out scenarios for future services. But EUDAT will also look for other topics that may lend itself for additional working groups. The new services survey is an important activity that may inspire us to contact experts in the respective areas, but EUDAT is also open to initiatives coming directly from communities.

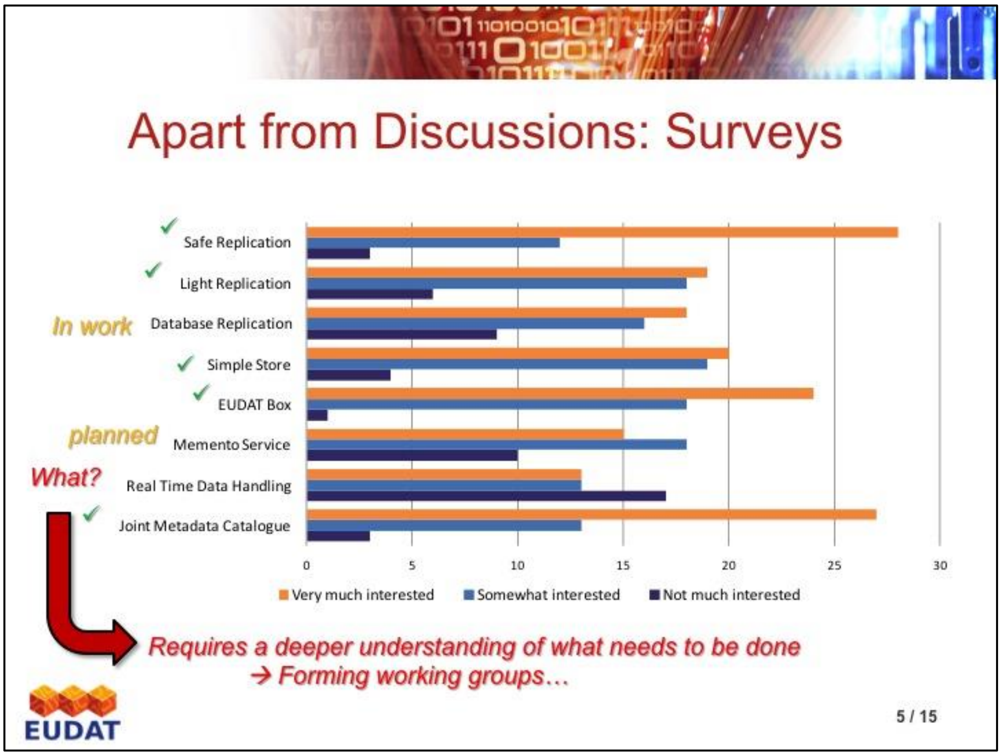
1.1.1 Listen to Community Requirements

Listening to communities and scientist requirements is core to the way EUDAT works and is not left for decision to an advisory board instead agile interaction is part of EUDAT's core discussion and decision bodies. It is important that the EUDAT Collaborative Data Infrastructure (CDI) provides building blocks and services, which are actually used and meet the needs of the communities and scientists. EUDAT adapted different ways to listen to communities and users and to involve field experts in defining the new building blocks and services. To identify new interest fields EUDAT engages directly with communities to discuss and identify specific requirements in the context of WP4 Stakeholder Requirements but also organizing surveys at EUDAT events (e.g. EUDAT User Forums and Conferences). Figure 1 shows preliminary results from a EUDAT Survey started in September and going on to the end of November[3].

[1] Dynamic data is data that is changing frequently without that humans control the changes such as in explicit versioning. One example of dynamic data are data streams generated by sensors but that have gaps due to technical reasons that are filled over time. Another example of dynamic data is given if you run experiments by massive crowdsourcing where you never know when exactly the participants will do certain tests.

[2] Data Access & Re-use is the subject of the fourth Working Group which was covered at the conference in Track 3

[3] The final results will be published on the EUDAT web-site.



Apart from Discussions: Surveys

Requires a deeper understanding of what needs to be done
→ Forming working groups…

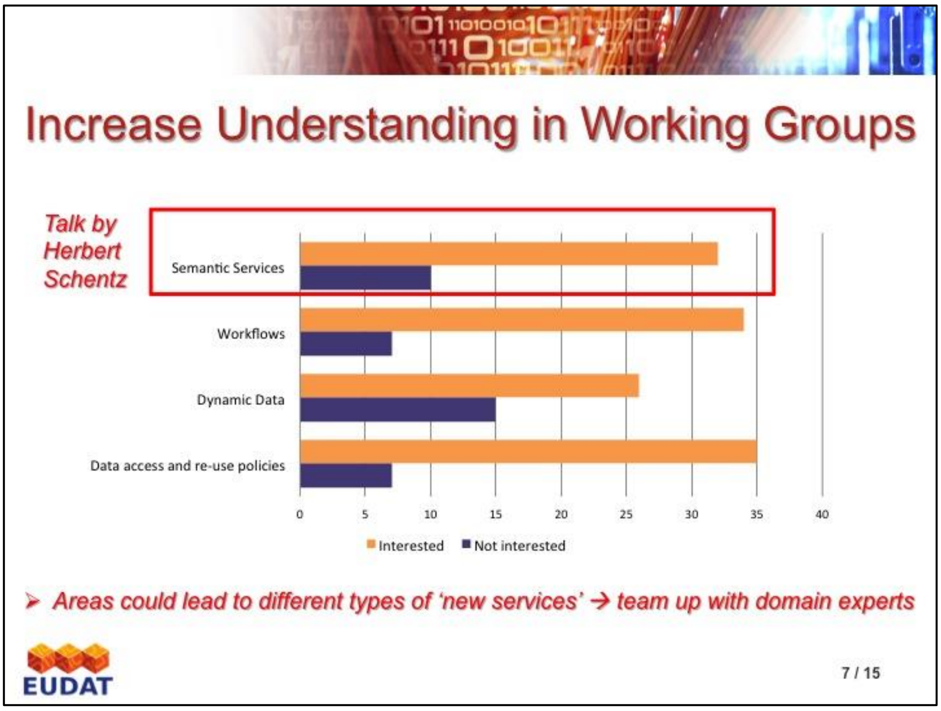
5 / 15

Figure 1 – New service Survey Result Overview example

When a specific new interest field is identified this is discussed in more technical detail in order to come to consensus on what a common service or building block is and what it should provide. To intensify this interaction in specific areas of interest EUDAT has adopted the concept of Working Groups from the DataONE[4] project. A working group is a method to bring domain experts, EUDAT community representatives and EUDAT technologists together to discuss identified interest fields, where the exact setup of a concrete service is not fully clear[5]. At the moment interest fields are identified on Semantic Annotation, Dynamic Data, Workflows and Data Access and Re-use Policies. Four working groups have been setup, one for each interest field, and domain experts were invited to join and to collaborate within a EUDAT working group. Figure 2 shows the relationship between a survey result and the working groups.

[4] https://www.dataone.org

[5] Also the FIM4R initiative in the area of federated identity management as started by Bob Jones and some colleagues can be compared with EUDAT's working group concept.



Increase Understanding in Working Groups

Talk by Herbert Schentz

➢ Areas could lead to different types of 'new services' → team up with domain experts

7 / 15

Figure 2 - Relationship between New Service Survey and Working Groups

To discuss the interest fields in detail in a Working Group, a workshop was held on the 25-26th September 2013 in Barcelona. One of the main conclusions from the EUDAT Working Groups workshop was that this concept is an excellent method to involve domain and service experts and therefore the established working groups will continue their collaboration. EUDAT is open to others to propose ideas for working groups where there is potential to identify common data services.

1.1.2 Semantic Annotation

Semantic Annotation is one of the new interest fields, which has been discussed in detail at the working group workshop and for which EUDAT is building a new building block. In this session the results from the Semantic Annotation track at the working group workshop; the LTER/LifeWatch use case, which has been the initiator on semantic annotation work in EUDAT; a new initiative on ontologies (EUON); and the work done on this subject in EUDAT were presented.

1.1.2.1 Semantic Annotation working group

Semantic Annotation is about connecting data with their meaning according to established ontologies or thesauri that are in general domain specific. One of the main barriers is to enable easy usage of these ontologies in the day-to-day working of a scientist. To do this, scientists need easy to use tools to bridge this gap or transparently make use of ontologies integrated with domain specific services. Solutions must be generic, lightweight, must follow semantic standards and must benefit from users semantic enrichments. The main conclusion of the working group meeting was that semantic annotation is of great interest to a broad user community and the working group is planning to continue its work.

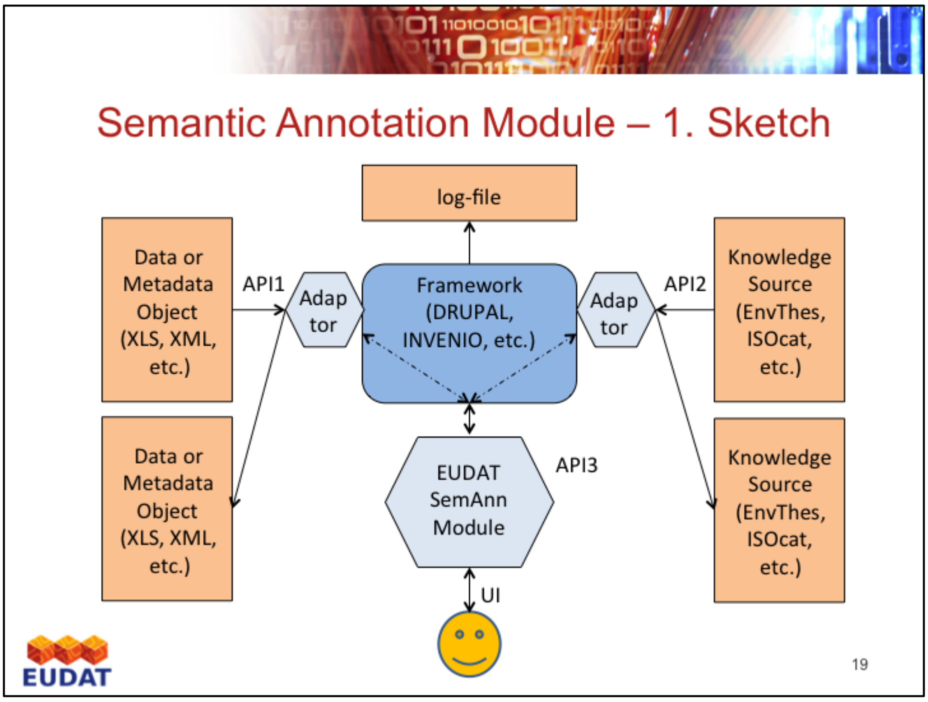
1.1.2.2 LTER/LifeWatch Semantic Annotation use case

The Europe Long-Term Eco system Research[6] (LTER-Europe) is part of the global International Long-Term Ecological Research[7] (ILTER). Part of the LTER objectives is to support cutting edge science with a unique in-situ infrastructure. The LTER/LifeWatch use case in EUDAT is to tackle the problem of the variety of data from the biodiversity domain and data from drivers of biodiversity, on existing and

[6] http://www.lter-europe.net/

[7] http://www.ilternet.edu/

gathered data in the field across 70 sites in Europe in compliance with existing metadata standards (e.g. EML and INSPIRE). The Semantic Annotation use case is trying to tackle the annotation of metadata with semantic concepts and terms and to discover data using semantic information. The goal of the Semantic Annotation service is to provide an easy-to-use tool which can be integrated within community frameworks, see Figure 3.



Semantic Annotation Module – 1. Sketch

log-file

Data or Metadata Object (XLS, XML, etc.)

API1

Adap tor

Framework (DRUPAL, INVENIO, etc.)

Adap tor

API2

Knowledge Source (EnvThes, ISOcat, etc.)

Data or Metadata Object (XLS, XML, etc.)

EUDAT SemAnn Module

API3

Knowledge Source (EnvThes, ISOcat, etc.)

UI

EUDAT

19

Figure 3 - Schematic overview EUDAT Semantic Annotation Module

European Ontology Network

During the session a new initiative on ontologies the European Ontology Network (EUON) was presented by Herbert Schentz (Umweltbundesamt GMBH). Large amounts of money went into the creation of domain specific ontologies, thesauri, vocabularies and frameworks, but these are hardly used. Why are these ontologies not used and what is missing? Practical tools and training for scientific data practitioners, easy access to semantic experts to support deployment of semantic web solutions, approaches to support the fast occurring changes and a platform for knowledge exchange between semantic experts and data practitioners. The goal of EUON is to connect the European Ontology Practitioner Community and to create a platform to provide quick and easy help to those who need to solve urgent problems in the semantic area. Membership is open and includes people from many scientific disciplines and from academic institutes, non-profit organisations and industry. EUON wishes to closely collaborate with EUDAT, to plan meetings at EUDAT conferences, user forums and working group meetings and specifically with EUDAT's Semantics working group.

1.1.2.3 Semantic Annotation Discussions & Conclusions

During and at the end of the Semantic Annotation session there were good discussions on the Semantic subject in general, on technical details and how the semantic annotation service is related to other EUDAT services. Remarks were made that the current solution is focusing on textual annotations and how this relates to non-textual data (e.g. pictures, audio, video); about the scalability of such a service and that a semantic annotation service is not the Holy Grail solving all semantic issues because scientists are the most knowledgeable in describing research and annotating is labour intensive. During the discussion about how the semantic annotation service is related to the other EUDAT services, options were discussed on the usage of the Semantic Annotation service within the B2SAFE, EPIC PID and B2SHARE services. The B2SHARE service will provide domain specific metadata templates to describe uploaded data objects.

The integration of the semantic annotation module within the B2SHARE service linked to domain specific vocabularies when selecting a domain specific template could be very beneficial to improve the quality of the metadata describing uploaded data objects. The main conclusions from this session are that any Semantic Annotation service should be generic, easy-to-use, scalable, flexible to handle different type of data objects and preferably, via auto learning technics, a high level of automation.

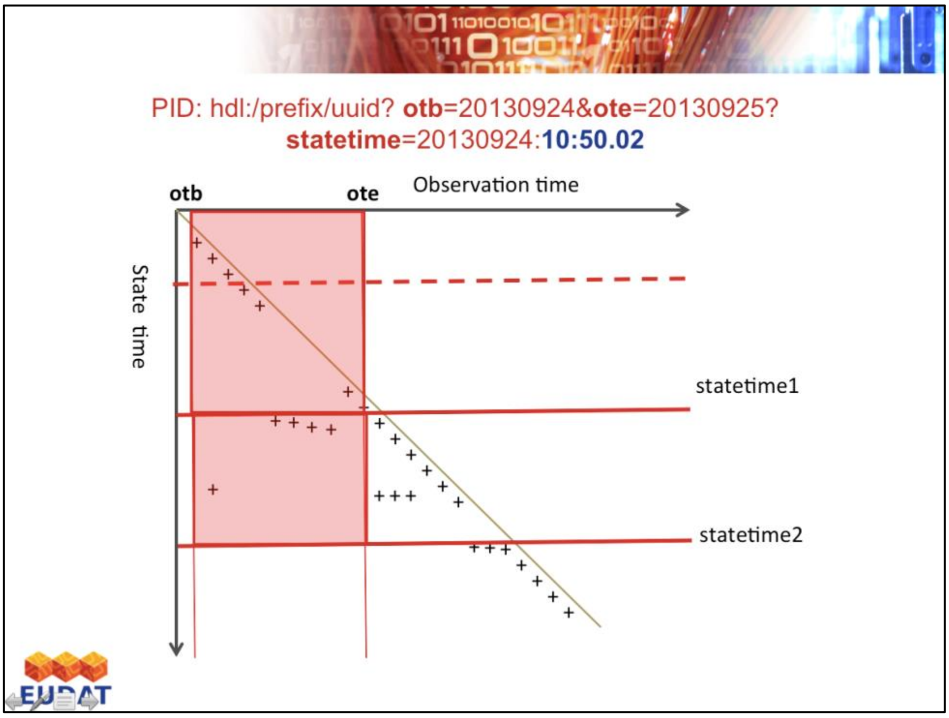
1.1.3 Dynamic Data

During the service building process and roll out of the Safe Replication (B2SAFE) service dynamic data has been a challenging subject. It is difficult to keep consistency between data objects, which are eligible to change and are replicated in a distributed environment. This use case is prominent within the seismology community (EPOS[8]) dealing with sensor-generated data in earthquake sensitive areas across Europe and data streams that are generated by mobile devices at unpredictable times and in unpredictable order (CLARIN[9]). Dynamic data is a broad subject, not only from sensor-generated data, but is seen within communities who have to deal with many unstructured and independent non-scientists (e.g. citizen scientists or crowdsourcing). Dynamic data is one of the working group interest fields. In this session the outcome of the Dynamic Data working group and the EPOS and CLARIN community presented their use case on dynamic data.

1.1.3.1 EPOS (European Plate Observing System) use case

The European Plate Observing System (EPOS) collaboration brings together the European seismology community to come to a common vision and approach to enable innovative multidisciplinary research to better understand the physical processes controlling earthquakes, volcanic eruptions, unrest episodes, tsunamis, tectonics and earth surface dynamics. The goal is to establish a long-term plan to facilitate the integrated use of data, models and facilities from existing, and new distributed research infrastructures. The EPOS community is dealing with data generated by sensors situated at earth sensitive locations across Europe. Data is transferred via phone lines or via satellite or radio links. These transmission methods have a level of uncertainty in which data fragments are not received in the right order, sensor frames can be delayed between minutes, hours or days or are never received. The challenge is to keep consistency between these changing data objects in a distributed environment and methods to enable reproducible science. An important aspect to enable reproducible science is to be able to track which version of a data object has been used to generate scientific results. With data objects, which are eligible, to change in time you need a method to identify a version of an object at time X and a method to identify a time frame within an object. This subject was discussed extensively within the Dynamic Data working group at the Barcelona meeting. The result from the working group meeting was to use a bi-temporal scheme to identify a version of a data object with two separate timelines: observation time and state time. The observation time indicates the time frame of the event/measurement, described with a begin and an end time. The state time describes the state or version of an object at a time. The relationship between observation and state time and the difference between versions of the selected observation time between “otb-ote” at state time 1 and 2 is explained in Figure 4.

[8] http://www.epos-eu.org/
[9] http://www.clarin.eu



PID: hdl:/prefix/uuid? **otb**=20130924&**ote**=20130925?
statetime=20130924:**10:50.02**

Figure 4 - Diagram describing the difference between observation and state time

1.1.3.2 CLARIN MPI-PL The Language Archive use case

The CLARIN MPI-PL *The Language Archive* use case looks from a different angle at the Dynamic Data challenge. The Language Archive is a unit of the Max Planck Institute for Psycholinguistics concerned with digital language resources and tools. It provides a large data archive holding resources on languages worldwide. The service is open to everyone to store good quality data. In the current age with mobile devices, data is easily generated. The challenge is to predict, store, manage and curate this vast growing data volume, to track intellectual property, to ensure data privacy and to engage a diverse growing user population (e.g. crowd sourcing). Engaging thousands of subjects in tests by using mobile devices means that data from participants will come in at unpredictable moments and at unpredictable order, nevertheless researchers want to start using the results for calculating evidences and obviously using them for publications. Similar to the case in EPOS, the concern is thus how to cite to a data matrix that is being filled at random uncontrolled moments. Also here observation and state time are different, since mobile devices could be off-line while an experiment is carried out etc.

1.1.3.3 Dynamic Data Discussions & Conclusions

The main discussions focused on the terminology used to identify versions of dynamic data objects and about the feasibility of supporting this within persistent identifier and repository systems, about how to handle the vast growing data volumes and about intellectual property rights.

The proposed bi-temporal scheme (e.g. observation and state time) for data objects appears to be a proper solution for tackling scientific reproducibility issues, and for data analysis carried out on real-time data. During the discussions some similarities were drawn with the spatial science domain to identify location areas.

Consensus on the terminology used to define the states is important to enable referencing and accessing data objects on basis of a bi-temporal scheme. It is recommended to interact with the RDA data citation working group.

In the discussion about the vast growing data volumes, the challenge is to manage and to store the data volumes and to be reservedly in destroying data. The current approach is to store all data and not to delete data objects, because the value of the data in the future is hard to predict.

In the crowdsourcing use case there was considerable discussion on how to track intellectual property. In general, IPR is handled via informed consent. But it is questionable if people providing data have a full understanding of the meaning and consequences of informed consent.

1.1.4 Workflows

The workflow session was the final session in the new services track and it is also one of the working group interest fields. Workflows are a joint research activity in EUDAT in which communities (e.g. ENES and CLARIN) are assessing solutions in which community workflows can make use of the EUDAT services. The results from the working group workshop were presented followed by the ENES and CLARIN use cases on workflows and a proposal for a generic execution framework (GEF).

1.1.4.1 Workflow working group workshop

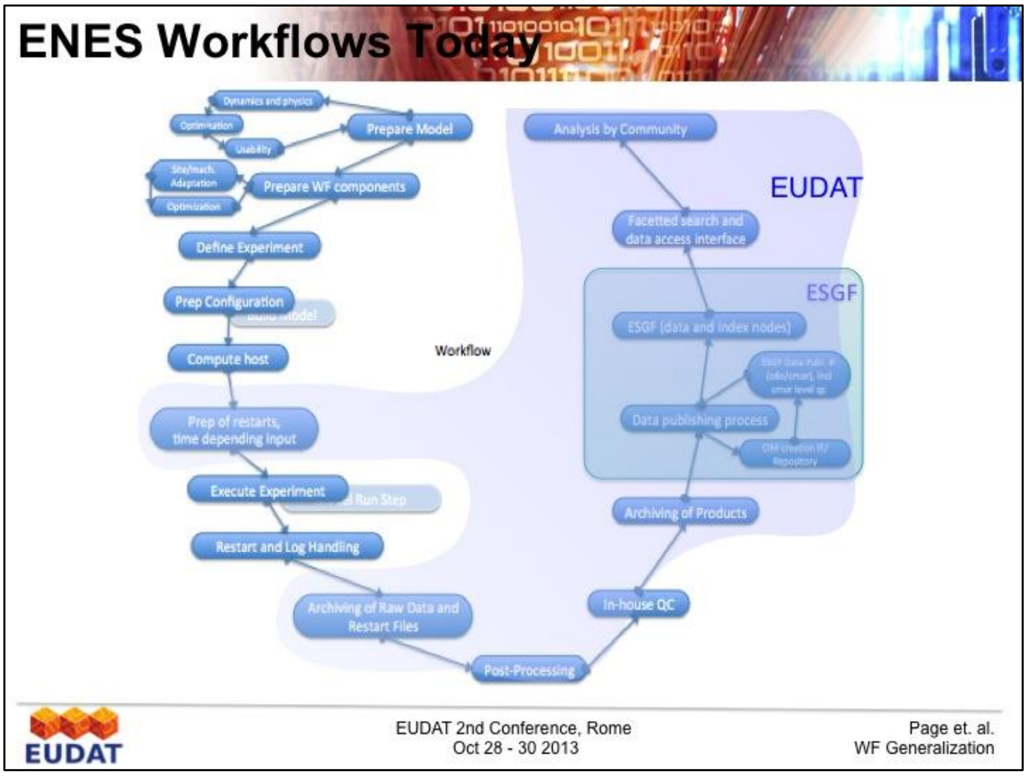
The goal of the working group workshop was to understand the needs of the community experts on common services, how to orchestrate data processing and how scientific workflows can make use of EUDAT services. Support for workflow provenance and services to register and describe workflow components and make them discoverable, referable (e.g. assigning PIDs to components) and to capture best practices were intensively discussed. It is very important to describe the functionality of a workflow component, input and output data formats and test data to certify the functionality of a component. Additionally to this it is recommended to EUDAT not to develop a new workflow system but rather to clearly define an API to be used within workflows. This is in line with the EUDAT GEF developments. The next steps are: not to lose momentum, to focus on concrete work and formalize and continue the work group.

1.1.4.2 ENES Workflow use case

The European Network for Earth System (ENES) community represents the European community for climate modelling providing predictions for the IPCC[10] report and for EU mitigation and adaptation of policies. Climate modelling studies are very compute intensive and there is a strong need for climate numerical models tailored for HPC computing. Therefore ENES is collaborating with PRACE to get access to world-class computing resources. Climate is a global event influencing all aspects of the environment. It impacts researchers and modellers from agriculture, water management, shipping, dikes, etc. EUDAT provides a platform and building blocks to enable this kind of inter-disciplinary research.

The ENES workflows must be integrated with the Earth System Grid Federation (ESGF), which is the workhorse of the ENES community. The workflow of tomorrow must provide better and more automated metadata management, provenance data, integrated into the daily work of the scientists and better interoperability between workflows in different communities. These were the main reasons to start working on workflows activity in EUDAT. Figure 5 gives a global overview of the relationship between the ENES workflows and the ESGF and EUDAT service domain.

[10] Intergovernmental Panel on Climate Change, homepage http://www.ipcc.ch



ENES Workflows Today

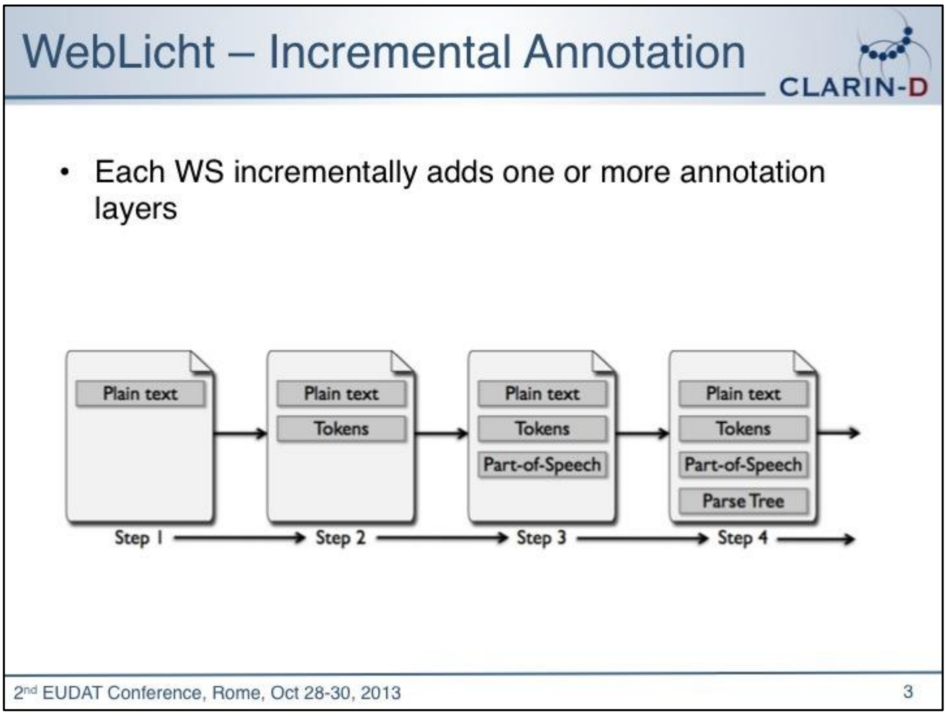
EUDAT 2nd Conference, Rome
Oct 28 - 30 2013

Page et. al.
WF Generalization

Figure 5 - Relationship between ENES workflows and ESGF and EUDAT service domain

1.1.4.3 CLARIN WebLicht workflow use case

WebLicht (Web-based Linguistic Chaining Tool) is a broadly used workflow engine for linguistic annotations, which is based on Service Oriented Architecture (SOA) principles. Each tool is made available as a web service. The user does not have to install any annotation tool on his/her local machine and is able to visualize the workflow within the web interface. In WebLicht each workflow step incrementally adds one or more annotation layers as shown in Figure 6. There are many challenges due to (1) increasing data sizes and (2) an increasing amount of users who want to execute chains. The increasing size and partly also legal issues of data make it hardly possible to move data to the locations where data analysis tools are being executed.. The increasing data volumes and the fact that an increasing amount of users want to execute workflow chains with their data require a change of approach so that data will be stored close to HPC or large cluster servers dependent on the type of algorithms being executed. The question is how WebLicht workflows can take advantage of EUDAT services which take care of storing data and bringing data close to computational facilities? A possible solution is to enrich the EUDAT B2STAGE to interface with workflows, for example with the generic execution framework.



WebLicht – Incremental Annotation

CLARIN-D

- Each WS incrementally adds one or more annotation layers

Plain text

Plain text
Tokens

Plain text
Tokens
Part-of-Speech

Plain text
Tokens
Part-of-Speech
Parse Tree

Step 1 → Step 2 → Step 3 → Step 4 →

2nd EUDAT Conference, Rome, Oct 28-30, 2013 3

Figure 6 - Web-Licht incremental annotation

1.1.4.4 EUDAT Generic Execution Framework

The idea of the Generic Execution Framework (GEF) is to enable processing of datasets close to where the data is stored, allowing faster access, lower network bandwidth usage and easier filtering and sub-setting by only transferring the end results back to the user. The GEF will provide an API layer, which consists of a collection of HTTP web services. The API layer allows easy integration of the GEF into existing workflow engines (e.g. Taverna, Kepler) and community specific data federation interfaces. GEF is built on top iRODS, which is the current core technology of the EUDAT Safe Replication (B2SAFE) service, but other back-ends are possible. It allows the input and output of data sets to be specified via URIs or handles/PIDs. The GEF API is generic, whereas functions are to be created and maintained by communities, functions can be combined into pipes. A pilot implementation of the GEF framework has been developed and has been tested within the ENES and CLARIN workflows. Tests are on the way to test a full integration within the ENES and CLARIN federations.

1.1.4.5 Workflow Discussion & Conclusions

The first questions were technical ones about the service implementation and the API functions supported by the GEF framework and how much training is needed to make use of the infrastructure. The service has been implemented in JAVA and provides an HTTP/Rest API interface. The basic functions are: send/get data and send/get workflows, which can only be executed by a community manager. On the subject of training an example was given about the uptake of the WebLicht workflow engine within the CLARIN community. This has been good and unexpected; currently the WebLicht workflow engine is used for teaching purposes at a number of universities to explain linguistics.

During the discussions comments on the results from the Workflow working group were given:

- EUDAT should try to minimise the number of workflows, but adopt a bottom up approach;

- EUDAT should look at cross community aspects and information about workflows should be discoverable. For this EUDAT could provision a workflow repository and registry service in which communities can provide content about workflow execution engines.